Paper ID: 282

Clustering Crimes Related Twitter Posts using WordNet and Agglomerative Algorithm

S.P.C.W Sandagiri, B.T.G.S Kumara, K. Banujan

Department of Computing and; Information Systems, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

bhakhuha@appsc.sab.ac.lk

Abstract: Crime is a major problem faced today by society. Crimes have affected the quality of life and economic growth badly. We can identify the crime patterns and predict the crimes by detecting and analyzing the historical data. We can use social media like twitter to detect crimes related activities. Because Twitter users sometimes convey messages related to his or her surrounding environment via twitter. In this paper, we proposed a machine learning approach to cluster the crime-related twitter post based on the crime category. The empirical study of our prototyping system has proved the effectiveness of our proposed clustering approach.

Keywords: Clustering, WordNet, Agglomerative algorithm, SVM

Introduction

A crime is a criminal act and can be enforced by a state or other body. Police collect criminality data from the field based on an individual or group's records. The police will determine the crimes which have occurred from the data. Police may not have evidence of violence from other groups (Gemasih et al., 2019). Research of crime has been a vital tool to help law enforcement officers protect people. Crime levels have risen significantly as a result of an increasingly growing population and effective research has been a time-consuming endeavor (Hissah and Al-Dossari, 2018).

We are now at a point where many human needs and requests can be found in online resources (Yang et al., 2011). In addition to the increasing market, there are also many officials, military, medical and private knowledge available online (Hernández et al., 2016). The proliferation of online infrastructure has also improved the way users connect on-line. Several web services allow users to communicate in real-time (Hernández et al., 2016).

Nowadays, with a rising number of Internet users, and the ease of connectivity provided by the proliferation of mobile data technologies, the amount of knowledge related crimes to be accessed and analyzed is growing correspondingly. Much of this material remains unstructured in the context of "free text." This phenomenon has resulted in a growing interest in the production of unstructured data processing approaches (Hissah and Al-Dossari, 2018). The material in social media provides rich qualitative knowledge about the everyday lives of its users, based on textual details shared. Each textual post is compiled on the service providers platform (Chen et al., 2015).

There are currently many social media sites, such as Facebook, Twitter, and Snapchat Twitter, which is one of the most common social networking sites for casual chats, sharing photos and ideas and transferring information and news via text, limited to 280 characters, called "tweets" (Hissah and Al-Dossari, 2018). Social networking sites can uncover useful information as a systematic

analysis of their unstructured data is implemented (Chen et al., 2015).

Text mining is an important method and can help to address this issue by efficiently classifying crimes. The program introduced for identifying and classifying crime related posts on Twitter (Hissah and Al-Dossari, 2018)

Throughout time, experiments have been undertaken to understand the nature of criminal behavior, classifying people by their racial or cultural context, or designing techniques of deterrence, detection, and 'effective' discipline. Governments around the world are investing huge amounts on crime prevention, law enforcement, and public intelligence. As a result, in recent years people have observed the growing number of CCTV cameras looking at the streets of all big cities, which is controversially detected and discussed. However, crime monitoring is now enabled by electronic networks and databases that include access to virtualized neighborhood observation and official records of crime (Bendler et al., 2014).

Although most experiments are restricted to particular places, forms of crime, neighborhoods, and consumers, or concentrate on unique incidents, we will generalize our suggested solution to any area. Hence, the primary aim of this work is to collect available, reliable information (tweets) to recognize the essence of crimes and to assist law enforcement with future crime reduction, thus adding to the welfare of mankind. As an initial step, we clustered the Crime Related Twitter Posts using WordNet and Agglomerative Algorithm.

The remainder of the paper is organized as follows. Section II describes the Literature Review. Section III explains the proposed approach while Section IV explains results and discussion. Finally, Section V concludes the paper with future directions.

Literature Review

Analysis of crime was researched extensively and different theories and methods emerged (Wang et al., 2012, Eck et al., 2005, Gerber, 2014, Caplan and Kennedy, 2011, Wang and Brown, 2011). Cohen and Felson concentrated primarily on offenses requiring overt interaction with criminals and individuals who were targeted (e.g. physical assault). They called these offenses malicious abuses of direct contact (Cohen and Felson, 1979).

The details that Twitter users share in a tweet normally include something specific to themselves or their environment, like the incident of a crime (Gemasih et al., 2019). GPS-tagged Twitter data allow for potential crime prediction in major cities (Chen et al., 2015)

Wang et al. (2012)'s previous research concentrated solely on recent department tweets to establish the correlation between subjects used in tweets and different forms of criminal accidents (Wang et al., 2012). Gerber (2014) further developed the prediction model by subjectmodeling (Gerber, 2014). He combined incidents of historical crime with GPS-tagged Twitter data collected from all twitter users in the Chicago city area. Such simulations, however, mostly found subject modeling but did not extend Twitter data to sentiment analysis.

Gonzalez et al. (2008) measure the distribution of human spatial probability by measuring regularities in both the temporal and spatial dimensions. The authors focus on their analysis of spatial data obtained from cell phone usage, where each user's estimated location can be determined from the mobile phone towers their phone is registered at. Arase et al. (2010) rely on consumer trip data to include recommendations for travel routes based on trends collected. Similarly, Scellato et al. (2011) forecast device positions based on

trends discovered by measuring data from a non-linear time series. The authors test their approach on various data sources (e.g. GPS tracks, WiFi connection points) and can substantially increase prediction precision over Markov Spatio-temporal predictors. Cho et al. (2011) indicate that the patterns of human migration are strongly intermittent, but contribute in part to the studied individual's social network when it comes to long-term travel. Backstrom et al. (2010) go further and prove that a person's geospatial location can only be predicted by individual locations within the respective social network.

Proposed Approach

Figure 1 shows the overall methodological framework used for the study. First, twitter posts are extracted using crime related keywords. Then, preprocessing techniques are applied to clean the data set. Next, twitter posts are transformed into vectors to generate the feature vectors in the data preparation step. Then, the Support Vector Machine (SVM) model is constructed to classify the data set. Next, calculate the similarity between twitter posts using WordNet. Finally, the agglomerative clustering algorithm is applied to classify the posts.

A. Data collection

Twitter posts are collected through the Search API (available at http://twitter.com) of Twitter. The search of the twitter posts must be based on a set of keywords that can be used to classify the crime situations. Then, twitter posts are labeled based on the contents used to create the training set. The collected data set consists of more than 100,000 twitter posts from 2020 January 01 to 2020 January 31.

B. Data Pre-processing

As a next step, pre-processing techniques should be applied to the extracted collection of data. Since typos, unnecessary items such as URLs and stop words in the twitter post can be available. Data gathered from twitter are often extremely unstructured and noisy. Clean tweet data are produced by pre-processing techniques that will be used for the next process.

First, we deleted stop words like, is, which, the, have, etc. The words convey no positive or negative significance. So, without affecting the meaning of the message, we can easily remove the stop word. We deleted then URLs, hashtags, symbols, usernames, terms, quotes, etc. Next, combine words with tokenization techniques. Finally, we used a stemming algorithm to simplify the word to a stem that contains suffixes, prefixes, or word roots.

Some tweets may contain text that seems irrelevant to the process of an analysis of sentiments. Candidate markers such as URL's, responses to other users and frequently occurring stop words are considered noise (Choy, 2012). To remove such unwanted contents, a noise reduction method was used.

C. Data Preparation

After completing the pre-processing, twitter posts are transformed into vectors to generate the feature vectors. The vectors are used in the learning phase for machine learning algorithms. In this research, we used the term frequency-inverse document frequency (TF–IDF) values to create the vectors. TF–IDF value reflects the importance of a term in a document to the collection of documents as in the following equation (Equation 1).

$$tfidf_{x,p} = tf_{x,p}\ log\ log\ \left(\frac{n}{pn_x}\right) \qquad (1)$$

Here, $tfidf_{x,}$ is the TF–IDF value for term x in post P. tf_{xp} is the term frequency of term x in post P. pn_x is the number of posts that contain term x. Parameter n is the total number of posts.

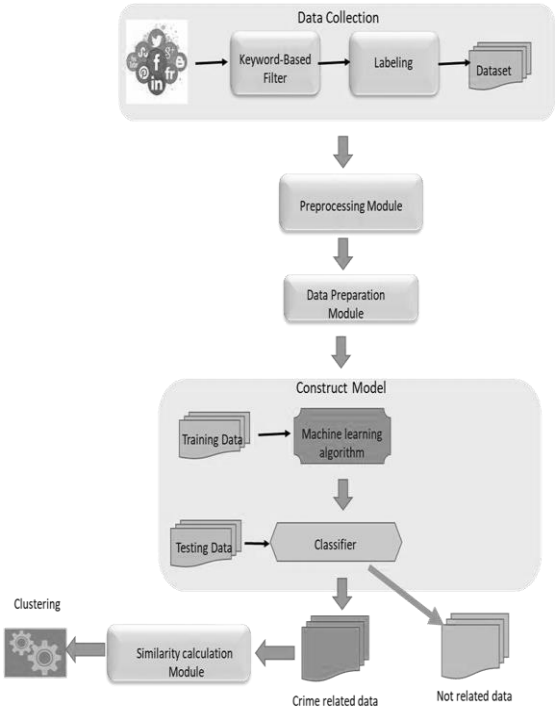

Figure 1: High-level architectural framework

D. Constructing the SVM model

We have used the SVM as the machine learning algorithm. The SVM is a state-of-the-art classification method that uses a learning algorithm based on structural risk minimization. The classifier can be used in many disciplines because of its high accuracy, ability to deal with high dimensions, and flexibility in modeling diverse sources of data. The output of SVM indicates the distance between testing data and the optimal hyperplane. We

prepared a training dataset D = [(x_1, y_1), (x_2, y_2),…, (x_n, y_n)], where x_i is the feature vector and y_i is the expected class label for the ith instance. The SVM is then trained with the labeled feature vectors to categorize the crime-related data and non-related data.

E. Calculating semantic similarity between twitter posts

After identifying the crime-related post using SVM, we need to cluster the post based on the crime types. As the first step in this stage, the semantic similarity between posts is calculated. We used WordNet as the knowledge base to calculate the semantic similarity of the posts.

WordNet is a lexical database of semantic relationships in more than 200 languages between words.WordNet connects words into semantic relationships that include synonyms, hyponyms, and meronyms. The synonyms are organized into synsets with descriptions of brief meanings and use.

WordNet resembles a thesaurus superficially, in that it brings together words based on their meanings. There are some major differences, however. Firstly, WordNet not only interlinks word forms - strings of letters - but unique words meaning. As a consequence, terms contained in the network close to each other are disambiguated semántically. Second, WordNet marks the semantic associations between words, while the groupings of terms in a thesaurus follow no other clear pattern than similarity.

Here to calculate the semantic similarity using WordNet, we applied an edge-count-based approach. First, the similarity between individual term pairs are calculated as follows (Equation 2);

$$Semantic\_(T_{ai}, T_{bj}) = WN\_Sim(T_{ai}, T_{2bj}) \quad (2)$$

Here, T_{ai} is the i^{th} term of post a, and T_{bi} is the j^{th} term of post b.

Then, the final similarity value between post a and b are obtained using the following equation (Equation 3).

$$Sim(a,b) = \sum_{p=1}^{l} \sum_{q=1}^{m} \frac{max\_sim(x_p, y_q)}{(l+m)} \quad (3)$$

Where x_p and y_q denote the individual terms, with l and m being the number of individual terms in twitter posts a and b, respectively.

F. Clustering crimes related posts.

We used an agglomerative clustering algorithm (Algorithm 1) as the clustering algorithm. This bottom-up hierarchical clustering method starts by assigning each twitter post to its cluster (see Line 1 in Algorithm 1). It then starts merging the most similar clusters, based on proximity of the clusters at each iteration, until the stopping criterion is met (e.g., number of clusters) (see Lines 4–10 in Algorithm 1).

Table 1: Algorithm 1 Clustering Algorithm.

Algorithm 1 Clustering	
Algorithm. **Input** S: Array of similarity values	
Input n: Number of required clusters	
Output C: clusters	
1:	Let each twitter posts be a cluster;
2:	ComputeProximityMatrix(S);
3:	k=NoOfPosts;
4:	**while** k !=n **do**
5:	Merge two closest clusters;
6:	k=getNoOfCurrentClusters();
7:	Calculate the center value of all posts in all clusters.
8:	Select post with the highest value of each cluster as cluster centers;
9:	UpdateProximityMatrix();
10:	**end while**

Results and Discussion

The experiments were conducted on a computer running

Microsoft Windows 10, with an Intel Core i5-3770, 2.70 GHz CPU and 6 GB RAM. Python was used for SVM implementation. Java was used as the programming language in implementing the similarity calculation module and the agglomerative clustering algorithm. Crimerelated tweets were collected from Twitter API.

A. Cluster Evaluation

We used precision, recall, and F-measure by using Equation 4, Equation 5, and Equation 6 respectively to measure the performance of our approach. Precision is the fraction of a cluster that comprises twitter posts of a specified class. A recall is the fraction of a cluster that comprises all posts of a specified class. We implemented a clustering approach using cosine similarities for comparison. Table 2 shows the experiment results.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$f_1 = 2.\frac{Precision.Recall}{Precision+Recall} \quad (6)$$

According to the experimental results (Table 2), for all other clusters, the WordNet-based approach obtained higher values for all the evaluation criteria. For example, the WordNet-based approach improved the precision value for the “Drugs Violations” cluster by 15.2%. Further, the WordNet-based approach improved the recall value for the “Assault” cluster by 16.0%.

B. Sample clustering results

Table 3 shows the sample outputs of each cluster.

Table 2 Performance measures of clusters

Conclusion and Future Works

In this paper, we proposed an approach to cluster the crimerelated posts from twitter. Here, first, we used the WordNet-based method to calculate the semantic similarity between twitter posts. Then, we applied the agglomerative clustering algorithm to cluster the posts based on calculated similarity values. The empirical study of our prototyping system has proved the effectiveness of our approach. It obtained 87.7% and 87.4 average precision and recall values respectively.

In future work, we planned to implement the crime prediction approach using SVM and deep neural network approach. Further, we planned to validate the tweets using news posts.

Cluster	WordNet-based Approach			Cosine similarity-based approach		
	Precision %	Recall %	F-Measure %	Precision %	Recall %	F-Measure%
Assault	86.9	86.0	86.4	70.0	75.3	72.6
Burglary	85.0	91.0	87.9	80.0	83.3	78.7
Drugs Violations	93.2	82.0	87.2	78.0	60.2	68.0
Homicide	83.8	93.0	88.2	81.0	74.3	77.5
Sex Offences	89.5	85.0	87.2	73.0	77.7	75.3

Table 3: Sample output

Cluster	Sample twitter posts
Assault	Police: Gun fires a shot after 6-year-old brings a gun to school in Wilcox County; parents arrested. , 26 year old Gaza fisherman moderately injured by gunfire after Egyptian boats opened fire on fishermen at sea , A high ranking source at APD just told me, The 27-year-old man in custody for the Violent knife attack at Freebirds in South Austin Friday morning is a homeless/transient with a violent criminal history. @fox7austin ,
Burglary	A foreign national was arrested, after being found in possession of a Toyota Corolla which was hijacked in Randburg. The driver reported the hijacking after he was taken hostage during the event., The New York police department arrested two teenage boys and charged them with gang assault and murder of 60-year-old Juan Fresnada that occurred during a $1 robbery on Christmas Eve., The truck used in the robbery was stolen from Greenville, South Carolina, where three of the four occupants were from.,
Drugs Violations	State Police Arrest Man for DWI and Seize $315 Worth of Drugs During Traffic Stop The #NJSP have arrested Carl Welch, 39, of Kenvil, N.J. and Michael Zelaya, 30, of Dover, N.J. and seized $315 worth of heroin and Xanax pills during a traffic stop, An Oklahoma City man was sentenced last month to 30 years in prison for drug and firearm possession connected to drug trafficking, Holyoke man arrested after police raid 2 homes, confiscate 2,500 bags of heroin , 1 lb. marijuana, gun.,
Homicide	A Dubuque man has been arrested in Minnesota on a vehicular homicide charge related to a fatal motorcycle crash in August. , Henrico police have made an arrest into Thursday nights homicide . Deion L. Smallwood, 23, of Richmond was arrested and charged with murder.
Sex Offences	A 55-year-old Sahuarita man was arrested on suspicion of sexual abuse Tuesday., Man sentenced to six years for north Edinburgh sex offences , CHILLING: Three people have been arrested in a child abuse investigation after Alabama authorities say children were locked in cages ,

References

ARASE, Y., XIE, X., HARA, T. & NISHIO, S. Mining people's trips from large scale geo-tagged photos. Proceedings of the 18th ACM international conference on Multimedia, 2010. 133-142.

BACKSTROM, L., SUN, E. & MARLOW, C. Find me if you can: improving geographical prediction with social and spatial proximity. Proceedings of the 19th international conference on World wide web, 2010. 61-70.

BENDLER, J., RATKU, A. & NEUMANN, D. 2014. Crime mapping through geo-spatial social media activity.

CAPLAN, J. M. & KENNEDY, L. W. 2011. Risk terrain modeling compendium. *Rutgers Center on Public Security*, *Newark*.

CHEN, X., CHO, Y. & JANG, S. Y. Crime prediction using twitter sentiment and weather. 2015 Systems and Information Engineering Design Symposium, 2015. IEEE, 63-68.

CHO, E., MYERS, S. A. & LESKOVEC, J. Friendship and mobility: user movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011. 1082-1090.

CHOY, M. 2012. Effective listings of function stop words for twitter. *arXiv preprint arXiv:1205.6396*.

COHEN, L. E. & FELSON, M. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review*, 588-608.

ECK, J., CHAINEY, S., CAMERON, J. & WILSON, R. 2005. Mapping crime: Understanding hotspots.

GEMASIH, H., RAYUWATI, R., SN, A. & MURSALIN, M. 2019. *Classification of Criminal Crimes From Data Twitter Using Class Association Rules Mining*.

GERBER, M. S. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125.

GONZALEZ, M. C., HIDALGO, C. A. & BARABASI, A.L. 2008. Understanding individual human mobility patterns. *nature*, 453, 779-782.

HERNÁNDEZ, A., SANCHEZ, V., SÁNCHEZ, G., PÉREZ, H., OLIVARES, J., TOSCANO, K., NAKANO, M. & MARTINEZ, V. Security attack prediction based on user sentiment analysis of Twitter data. 2016 IEEE international conference on industrial technology (ICIT), 2016. IEEE, 610617.

HISSAH, A.-S. & AL-DOSSARI, H. 2018. Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 9.

SCELLATO, S., MUSOLESI, M., MASCOLO, C., LATORA, V. & CAMPBELL, A. T. Nextplace: a spatio-temporal prediction framework for pervasive systems. International Conference on Pervasive Computing, 2011. Springer, 152-169.

WANG, X. & BROWN, D. E. The spatio-temporal generalized additive model for criminal incidents. Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, 2011. IEEE, 42-47.

WANG, X., GERBER, M. S. & BROWN, D. E. Automatic crime prediction using events extracted from twitter posts. International conference on social computing, behavioral-cultural modeling, and prediction, 2012. Springer, 231-238.

YANG, Y., LI, H. & DENG, G. A case study: behavior study of chinese users on the internet and mobile internet.International Conference on Internationalization, Design and Global Development, 2011. Springer, 585-593.

Author Biographies

S. P. C. W Sandagiri an undergraduate student in the Sabaragamuwa University of Sri Lanka and will be graduating in 2020 with a BSc Special in Information Systems.

Banage T. G. S. Kumara received the Bachelor's degree in 2006 from Sabaragamuwa University of Sri Lanka. He received the master's degree in 2010 from the University of Peradeniya and the Ph.D. degree in 2015 from the University of Aizu, Japan. His research interests include semantic web, web data mining, web service discovery, and composition.

Kuhaneswaran Banujan received his Bachelor of Science degree in 2019 with the Second Class Upper Division from Sabaragamuwa University of Sri Lanka. He is currently attached to the Department of Computing and Information Systems as a Lecturer in Computer Science. His research interests include Data Mining, Knowledge Management, Ontology Modeling, Business Process Simulation